Proposed NSLRSDA Data Sieve:
Assessment and Recommendations

Introduction

The National Satellite Land Remote Sensing Data Archive (NSLRSDA) was established by Public Law 102-555 *to maintain an archive of land remote sensing data for historical, scientific, and technical purposes, including long-term global environmental monitoring.* The responsibility of the NSLRSDA lies with the U.S. Department of Interior, specifically the U.S. Geological Survey's EROS Data Center (EDC).

One of the challenging tasks of the EDC archivist is to determine which data should or **should not** be accepted into the archive. The scope of this problem is illustrated in Figure 1. Less than five years ago the National Research Council[1] predicted a steady increase in the demand for archiving land satellite data. However, in reality, the archival demand is growing much faster than was expected. The slow start up reveals the delayed implementation of NASA Earth Observing System capabilities. However, by 2005, the demand for archival services will be at least one-third larger than expected in 1997, which suggests that the actual situation in 2005 could be far worse. At this rate of growth, the archive will not be able to meet the demand, both in terms of storage capacity and cost. What was projected to be near linear growth is highly non-linear. NSLRSDA needs to develop and implement data sieve methods to ensure that data most relevant to the archive's mission and directive are archived, thereby reducing the rate of data volume increases in the future. One uncertainty is the degree to which digital data handling and storage costs will improve. Current trends in technology development could make archiving large volumes of data a much less daunting task.

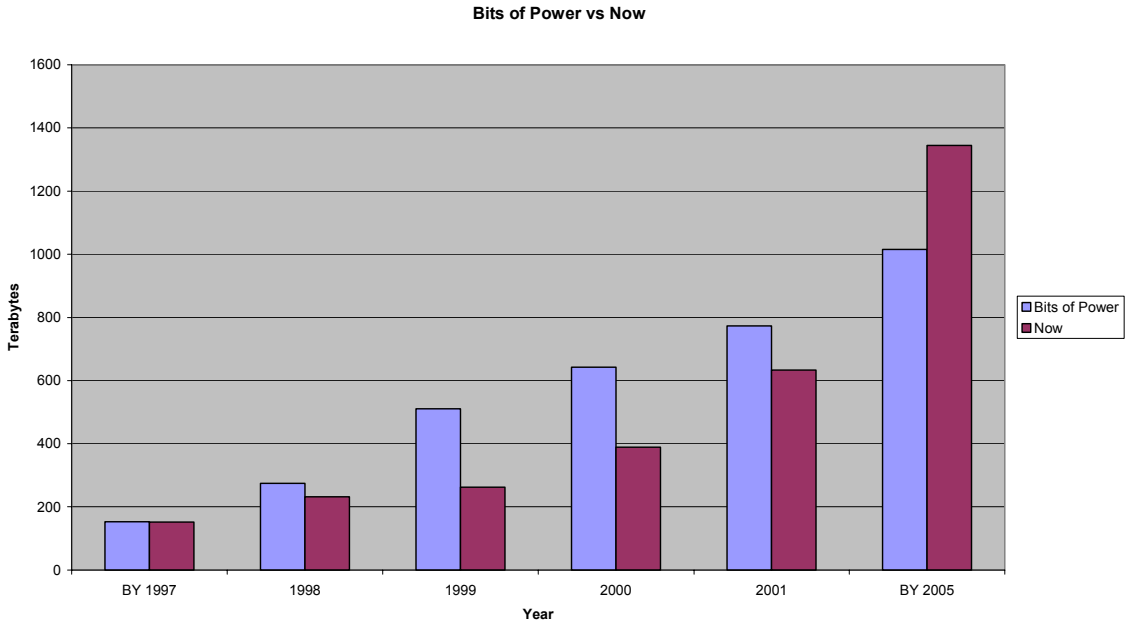

Figure 1. Archival demand over time from *Bits of Power* (National Research Council, 1997, versus current demand and projection from J. Faundeen, EDC Archivist.

[1] *Bits of Power*, National Research Council, 1997.

Assessment

Development of the data sieve is not an easy task. The initial reaction is that all data are important and therefore must be archived. But in all practicality, this approach is not feasible. Several attempts have been made to develop a strategy for assessing which data will be archived in NSLRSDA.

NSLRSDA Advisory Committee Considerations

Members of the first charter of the Advisory Committee to NSLRSDA (1998-2000) began development of a data sieve by defining a *Basic Data Set*[2] (Appendix A) in which priorities and criteria for considering candidate data were determined. Table 1 summarizes attributes of the *Basic Data Set Definition*.

Table 1. Summary of Basic Data Set Definition

Priorities	Criteria	Exclusions
Over US Land Territory from any civil (public/private) satellite	Data that provide an historical record of the land surface (base data for comparative analysis)	National Technical Means data anywhere
Over land anywhere from unclassified US satellite acquisitions	Data that relate to environmental global change monitoring and research	Oceanic and atmospheric data
Data collected by foreign remote sensing space systems, as determined to be appropriate by the Secretary	Data that are within the fiscal constraints of the archive	
	Data that establish processes and mechanisms to identify user information needs	
	Data that preserve datasets that are in risk of permanent loss	

When NSRLSDA implemented these guidelines **all** candidate data sets made it through the sieve. Nevertheless, this sieve serves as the foundation for further refinement and development of a set of recommendations to identify prospective data for inclusion in NSLRSDA. Other potential factors that might enhance the capacity of the archive to achieve its goals should be considered. This challenge was tasked to members of the second charter of the Advisory Committee (2000-2002). The primary tasks of the subcommittee during the second charter have focused on refining the basic data set definition and developing a set of sieve recommendations beyond that, which already have been considered by the internal USGS Data and Archive Review Team (DART).

Internal USGS Considerations

The USGS convened the Data and Archive Review Team (DART) to build upon the Basic Data Set. The team considered further technical and administrative details and proposed more detailed sieve criteria for NSLRSDA's use (Table 2). Tests of these proposed criteria by DART revealed that they actually did very little to reduce the volume of data sets that would be acceptable for inclusion in NSLRSDA.

The DART considerations are further amplified by the USGS technical considerations given to data sets being considered for possible inclusion in long-term archives (Appendix B). Although such constraints help to address some of the desirable technical characteristics of data that might be included in NSLRSDA, they do not provide guidance concerning what should be the conceptual goals of NSLRSDA.

[2] See *Basic Data Set Definition* on the NSLRSDA web site at http://edc.usgs.gov/archive/nslrsda/advisory/Definitions.html

Table 2
The USGS Data and Archive Review Team (DART)
Data Sieve Recommendations

CATEGORY	HIGH RANK	MEDIUM RANK	LOW RANK
Scientific Relevance	Well documented	Some known value	Unknown value
Data Volume	Low (1-10 TB)	Medium (11-50 TB)	Large >50 TB
Metadata	Complete	Partial	Absent
Distribution	Full and Open	Limited	None
Political Relevance	High	Medium	Low
Annual Budget	<$100K	$100K-$500K	>$500K
Funding	75%-100%	50%-75%	25-50%
Spatial	Global	National	Regional
Temporal	Older 20 years	10-20 years	<10 years
Span	>10 years	5 to 10 years	1 to 5 years
Risk of Loss	High	Medium	Low
Deadline to Decide	Now	1-2 years	>2years

External (User Community) Considerations
Within the International science community there has been substantial discussion of access to data, and expected archival practices employed, to facilitate scientific exploitation of terrestrial measurements. A particularly good summary of this perspective is provided in the National Research Council report *Bits of Power* (http://www.nap.edu/books/0309056357/html/index.htm). Within this report a set of principles concerning archival access to scientific data sets is presented, referred to as the "Bromley Principles" (Appendix C). These principles provide some useful insights on the science community perspective on what should be preserved and how access should be provided to such archived measurements. A key element of this report, related to the data sieve, is the preservation of data for long-term global change research. This emphasis also is placed on NSLRSDA under the basic data set definition.

A more practical example of how data sets might be evaluated relative to NSLRSDA is the long-term acquisition plan developed for the Landsat 7 mission operations (Arvidson et al., 2001). This operational framework was developed on the basis of scientific goals that the Landsat mission was designed to achieve. This is the first time in the Landsat mission history that such formal operations design has been included. The goal of this system is a quarterly, seasonal archive of clear land surface views for all regions of the globe. Such a framework, adopted (adapted?) to the archival needs of NSLRSDA might in fact provide a useful framework to further refine the data sieve considerations of the archive.

Recommendations

Despite the various efforts to selectively determine the types of data sets that will be considered for inclusion in NSLRSDA, the volumes of data being proposed for the archive continue to increase significantly. It seems unlikely that adequate fiscal resources will be made available to NSLRSDA to effectively address all demands that are placed on it. The Data Sieve subcommittee therefore recommends the following additional considerations to evaluate whether specific data sets should be considered for inclusion in the NSLRSDA.

- **Long-Term Observations**

Collecting observations of the Earth's land areas from space now has been carried out for nearly a half century. There are many sources of these observations, extending from dedicated observatories, such as Landsat and SPOT, to short-term experimental missions such as the original Skylab space station and the current NASA EO-1 mission. Although all such observations deserve some concern relative to preservation, **this committee recommends that high priority consideration be given to the long-term observations that provide consistent, repetitive coverage over extended periods of time.** The best example is the Landsat mission, which now has continuously monitored the Earth's land areas for over 30 years.

- **Special Collections**

As satellite remote sensing technology evolves there are many experimental activities that have been executed to test and evaluate new technologies. The NASA Earth Observer One mission, currently in orbit, is a good example. There have been many such activities since the early 1960s. **This committee recommends that experimental data sets be viewed as "special collections" primarily for historical interest with less ongoing science value than long-term systematic measurements.** As a result, in general, fewer resources would be dedicated to access and preserve these data sets.

- **Observations Focus**

There are a wide range of possible satellite Earth observatories that could be considered for inclusion in the NSLRSDA. These range from Geostationary orbits (and beyond) to low orbit, very fine spatial resolution, commercial imaging systems. Today the observations derived from these various systems can be categorized in terms of spatial and temporal resolution, which tend to inversely vary between differing sensors (Table 3).

Table 3
Spatio-Temporal Characteristics of Satellite Earth Observatories

Spatial Resolution	Temporal Resolution	Typical Sensor/Satellites	Typical Data Sources
10km – 1km	Daily -Hourly	NOAA Geostationary/Polar	NOAA/ESA Met Offices
1km – 100m	Monthly – Daily	AVHRR/MODIS/NPP/VIIRS	NOAA/NASA/DOD
100m – 10m	Quarterly – Weekly	Landsat/SPOT/ ASTER/ IRS	NASA/commercial
10m - <1m	5 years – Annually	Ikonos/Quickbird/EROS-A	Commercial

It is useful to note that the first row of data sets is well managed by NOAA and that the last row of observations are managed by the commercial industry exclusively (with the exception of the early "spy satellites such as Corona and the KH-series). This leaves the moderate resolution systems (e.g. Landsat/MODIS), which, interestingly enough are primarily dedicated to scientific monitoring of the Earth's land areas, as the primary archival goal of the NSLRSDA. **This committee recommends that NLSRSDA focus on compiling and making available the long-term, global records of land observations from the moderate resolution observatories.**

The committee further recommends, that NLSRSDA strives to acquire, for any given time-period, the "best" observations available for any given land location. "Best" in this regard should include:

- Lowest cloud cover possible
- Combination of spectral, spatial and radiometric precision
- Highest temporal frequency possible.

As a starting point, the Committee recommends that NLSRSDA should set a goal to acquire and maintain full global observation records as noted in Table 4

Table 4

Spatial Resolution	Temporal Resolution	Sources	Record Length
1km to 100m	Data to produce 10-day Cloud-Free Composites	AVHRR/MODIS/NPP/VIRRS	1982 - present
100m to 10m	Quarterly Cloud-Free (including data adequate to produce quarterly composites as needed)	Landsat, SPOT, ASTER, IRS	1972 - present

Because these moderate resolution observations, have come from multiple platforms and sensors (e.g. NOAA-7 to NOAA-1?, Landsats 1-7, RBV, MSS, TM, ETM+**), The Committee further recommends that NLSRSDA should work with NOAA and NASA to develop and maintain adequate information for cross-comparison of measurements from each of these sensors and platforms**. This information should include:

- Sensor –specific technical characteristics such as spectral band passes
- Time-dependent sensor calibration information
- Orbital characteristics such as timing and geographic spacing.

Appendix A

Basic Data Set Definition
(from archive web site)

Consistent with P.L. 102-555, the following priorities are the suggested focus for data acquisition by the National Satellite Land Remote Sensing Data Archive (NSLRSDA):

Priorities:

1) Over US Land Territory from any civil (public/private) satellite.
2) Over land anywhere from unclassified US satellite acquisitions
3) Data collected by foreign remote sensing space systems, as determined to be appropriate by the Secretary.

Criteria - data that:

- Provide an historical record of the land surface (base data for comparative analysis);
- Relate to environmental global change monitoring and research;
- Are within the fiscal constraints of the archive;
- Establish processes and mechanisms to identify user information needs.
- Preserve datasets that are in risk of permanent loss.

Exclusions:

- NTM data anywhere;
- Oceanic and atmospheric data.

Technical Characteristics of archived data:

- Archive raw imagery and ancillary data (radiometric calibration, atmospheric correction, motion model, etc. in standard formats)
- Stored by strip/pass rather than scene;
- Maintain a high quality ground control archive, distributable in standard formats wherever possible;
- Maintain DEM archives, distributable in standard formats wherever possible;
- Documented in catalogs accessible via browse facilities;
- Archive higher level products where appropriate (including ancillary data and standard metadata);

Technical Characteristics of retrieved data products:

- Retrieval by specified areas;
- Include geometry model, spacecraft motion model, etc.;
- Precision and terrain correction services available;

Appendix B

U.S. Geological Survey Data Transfer Requirements
For Long-Term Archiving

INTRODUCTION

This document is intended to ensure that the USGS can adequately address the archiving, access, distribution, and user service responsibilities for any data offered to the USGS. These requirements assume that the data have already met the scientific criteria relevant to the land science mission of the USGS. There may be some additional specific requirements related to data access, distribution, and user services that would entail further discussions. Lastly, the levels of service the USGS provides for new data are negotiable and will be handled on a case-by-case manner.

PHYSICAL REQUIREMENTS

- Data provided will be at the level of processing that best preserves the integrity of the data and is the most useful to the anticipated requestors of the data. USGS, in cooperation with the party offering the data, will determine this level. Ancillary files required for processing higher-level products are to be provided as well.
- The data shall be uncompressed, or, if the USGS agrees, utilize a loss-less compression technique accompanied with the decompression algorithm and any additional software needed to read or decompress the data.
- The data, if transferred on media, shall reside on 'state of the art' media (e.g. in Fiscal Year 2002, DLT 7000, 9940A, or 9940B) that is compatible with USGS systems and has at least five years of reliable life remaining.
- The format that the data is transferred in should be non-proprietary and computer-compatible. If the data format is proprietary, a sunset date when the data would be considered Public Domain shall be negotiated and agreed upon prior to any transfer.
- The data's file naming convention will be documented and provided to the USGS.
- For all raster data sets, the transfer shall include a non-proprietary raw, raster-formatted browse for each record or granule of data along with the documentation of how the browse was created.
- The initial, and any subsequent, processing histories must accompany the data. The processing histories should list the hardware and software environment at the time the processing was performed.
- Regarding analog aerial photography, the USGS will only accept camera original film on a polyester base stored in rubber cans. The manufacturer number of the film must also be provided. The photographic frame reference numbers must be on the film. No paper photography will be accepted.

METADATA REQUIREMENTS

- Complete metadata must be provided prior to any physical shipment of data and must be capable of supporting FGDC/ISO collection- and record-level standards with emphasis upon complete frame/image center and corner latitude and longitude coordinates.
- Metadata must be provided via flat, ASCII, delimited files and indexed to tie to the physical inventory of records being considered for transfer.
- Additionally, the data shall be accompanied by a complete library of documentation, including but not limited to user information (e.g. guides, DIFs, fact sheets, FAQ sheets, user guides, etc), instrument documentation (e.g. PDR and CDR information, lessons learned, hardware documentation, firmware documentation, engineering models, computer models, etc), platform documentation (e.g. overview, etc), algorithm documentation (e.g. ATBDs, 'grey' books, etc), and so on.
- The data's documentation shall be provided in hardcopy form where possible, and in a current softcopy format (e.g. in Fiscal Year 2002, Microsoft Word).
- The data provider will also provide a comprehensive training course in the science and prior management behind the instrument, data, processing histories, format(s), algorithms, and user community to the USGS science, engineering, and user services staff.

POLICY REQUIREMENTS

- Once a physical transfer of any data has occurred and the USGS has formerly accepted the data, the data become the responsibility and property of the USGS.
- Contributed data may be transferred to other data centers such as the National Archives and Record Administration when deemed appropriate by the USGS.
- Product or derivative product pricing will be determined by USGS guidelines including the elements that make up the principle of the Cost Of Fulfilling a User Request (COFUR).
- The data should have no access restrictions applied to them including any use policies that would prohibit unlimited distribution capabilities. If restrictions do apply to the data, a mutually agreed to formal sunset date for such restrictions to be removed and the data considered to be part of the Public Domain must be in place prior to any data transfers.
- A technical point-of-contact must be provided that can be utilized for questions about the data or use of the data. A USGS point-of-contact will also be established representing the USGS science, user service and the long-term archive areas.
- Requests to reprocess or to make large copies of data already submitted and accepted into the USGS long-term archive will be handled on a case-by-case basis and at the discretion of the USGS.

Appendix C

The "Bromley Principles" Regarding Full and Open Access to "Global Change" Data

The overall purpose of these policy statements is to facilitate full and open access to quality data for global change research. They were prepared in consonance with the goal of the U.S. Global Change Research Program and represent the U.S. government's position on access to global change research data.

- The Global Change Research Program requires an early and continuing commitment to the establishment, maintenance, validation, description, accessibility, and distribution of high-quality, long-term data sets.
- Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective.
- Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive. Procedures and criteria for setting priorities for data acquisition, retention, and purging should be developed by participating agencies, both nationally and internationally. A clearinghouse process should be established to prevent the purging and loss of important data sets.
- Data archives must include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.
- National and international standards should be used to the greatest extent possible for media and for processing and communication of global data sets.
- Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.
- For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they become widely useful. In each case, the funding agency should explicitly define the duration of any exclusive use period.

REFERENCES

Arvidson, T., Gasch, J., & Goward, S.N. (2001) Landsat 7's Long Term Acquisition Plan - An Innovative Approach to Building a Global Archive, Special Issue on Landsat 7. *Remote Sensing of Environment*, 78, 13-26.

National Research Council (1997). *Bits of Power: Issues in Global Access to Scientific Data*. National Academy Press, Washington, DC, 1997.
Also include as reference:

1. PL 102-555…